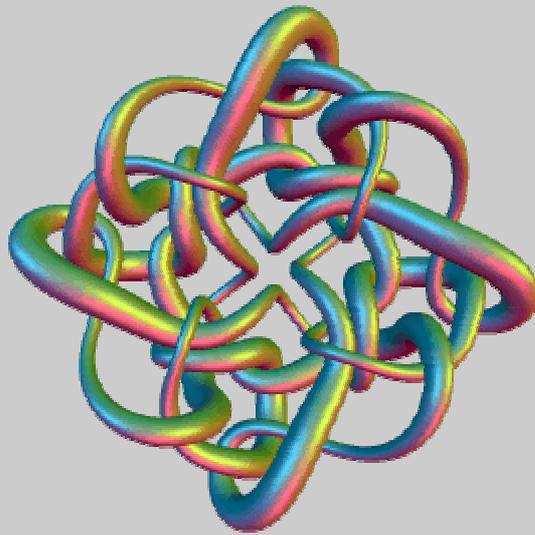
Material Didático

# Série
# Estatística Básica

## Texto III

# Amostragem & Estimação

Enfoque: Exatas

Prof. Lorí Viali, Dr.

# SUMÁRIO

# 1. AMOSTRAGEM

## 1.1. CONCEITOS BÁSICOS

**Estatística Indutiva.** Muitas vezes, apesar dos recursos computacionais e da boa vontade não é possível estudar todo um conjunto de dados de interesse. Neste caso estuda-se uma parte do conjunto. O principal motivo para se trabalhar com uma parte do conjunto ao invés do conjunto inteiro é o custo.

O conjunto de todos os elementos que se deseja estudar é denominado de *população*. Note-se que o termo população é usado num sentido amplo e não significa, em geral, conjunto de pessoas. Pode-se definir uma **população** como sendo:

**Uma coleção de todos os possíveis elementos, objetos ou medidas de interesse.**

Assim, são exemplos de populações:

- O conjunto das rendas de todos os habitantes de Porto Alegre;
- O conjunto de todas as notas dos alunos de Estatística;
- O conjunto das alturas de todos os alunos da Universidade; etc.

Um levantamento efetuado sobre toda uma população é denominado de levantamento censitário ou **Censo.**

Fazer levantamentos, estudos, pesquisas, sobre toda uma população (censo) é, em geral, muito difícil. Isto se deve à vários fatores. O principal é o custo. Um censo custa muito caro e demanda um tempo considerável para ser realizado. Assim, normalmente, se trabalha com partes da população denominadas de *amostras*. Uma **amostra** pode ser caracterizada como:

**Uma porção ou parte de uma população de interesse.**

Utilizar amostras para se ter conhecimento sobre populações é realizado intensamente na Agricultura, Política, Negócios, Marketing, Governo, etc., como se pode ver pelos seguintes exemplos:

- Antes da eleição diversos órgãos de pesquisa e imprensa ouvem um conjunto selecionado de eleitores para ter uma idéia do desempenho dos vários candidatos nas futuras eleições.

- Uma empresa metal-mecânica toma uma amostra do produto fabricado em intervalos de tempo especificados para verificar se o processo está sob controle e evitar a fabricação de itens defeituosos.

- O IBGE faz levantamentos periódicos sobre emprego, desemprego, inflação, etc.

- Redes de rádio e TV se utilizam constantemente dos índices de popularidade dos programas para fixar valores da propaganda ou então modificar ou eliminar programas com audiência insatisfatória.

- Biólogos marcam pássaros, peixes, etc. para tentar prever e estudar seus hábitos.

O processo de escolha de uma amostra da população é denominado de **amostragem**.

**Riscos da amostragem**. O processo de amostragem envolve riscos, pois toma-se decisões sobre toda a população com base em apenas uma parte dela. A **teoria da probabilidade** pode ser utilizada para fornecer uma idéia do risco envolvido, ou seja, do erro cometido ao utilizar uma amostra ao in-

vés de toda a população, desde que, é claro, a amostra seja selecionada através de critérios probabilísticos, isto é, ao acaso.

Baseado nos conceitos anteriores pode-se definir **Estatística Indutiva** ou **Inferencial** como sendo:

A coleção de métodos e técnicas utilizados para se estudar uma população baseados em **amostras probabilísticas** desta mesma população.

**Uma amostra é dita probabilística se todos os elementos da população tiverem probabilidade conhecida e não zero de pertencer a amostra.**

Dentre as várias maneiras de se selecionar uma amostra probabilística ou aleatória de uma população a mais simples é atribuir a todos os elementos da população a mesma probabilidade de pertencer a amostra.

Uma amostra que satisfaça tal critério é denominada de **amostra aleatória simples (aas).**

Uma aas pode ser extraída de uma população de acordo com os critérios:

**(a)** com reposição e **(b)** sem reposição.

Se a população for infinita então as retiradas com e sem reposição serão equivalentes, isto é, se a população for infinita (ou então muito grande), o fato de se recolocar o elemento retirado de volta na população. não vai afetar em nada a probabilidade de extração do elemento seguinte.

Se, no entanto, a população for finita (e pequena) será necessário fazer uma distinção entre os dois procedimentos, pois na extração com reposição as diversas retiradas serão independentes, mas no processo sem reposição haverá dependência entre as retiradas, isto é, o fato de não recolocar o elemento retirado afeta a probabilidade do elemento seguinte ser retirado. A amostragem sem reposição é mais eficiente que a amostragem com reposição e reduz a variabilidade uma vez que não é possível retirar elementos extremos mais do que uma vez.

Assim se **N** representa o tamanho da população e **n** < N o tamanho da amostra, então o número de amostras possíveis de acordo com os critérios com e sem reposição será:

## (a) Com reposição

k = número de amostras = $\mathbf{N^n}$

## (b) Sem reposição

k = número de amostras = $\binom{N}{n} = \frac{N!}{n!(N-n)!}$

<u>Exemplo:</u>

Considere a população P = { 1, 3, 5, 6 }. Então o número de amostras possíveis de tamanhos n = 2 e n = 3, de acordo com os critérios com e sem reposição será:

## (a) Sem reposição

**(1)** n = 2

Como N = 4 e n = 2, então o número de amostras possíveis será: $\binom{N}{n} = \binom{4}{2} = \frac{4!}{2!(4-2)!} = 6$

Estas amostras serão: (1, 3)  (1, 5)  (1, 6)  (3, 5)  (3, 6)  (5, 6)

**(2)** n = 3

Como N = 4 e n = 3, então o número de amostras possíveis será: $\binom{N}{n} = \binom{4}{3} = \frac{4!}{3!(4-3)!} = 4$

Estas amostras serão: (1, 3, 5)  (1, 3, 6)  (1, 5, 6)  (3, 5, 6).

## (b) Com reposição

**(1)** n = 2

Como N = 4 e n = 2, então o número de amostras possíveis será $N^n = 4^2 = 16$.

Estas amostras serão: (1, 1) (1, 3) (1, 5) (1, 6) (3, 3) (3, 5) (3, 6) (5, 5)
(5, 6) (6, 6) (3, 1) (5, 1) (6, 1) ( 5, 3) (6, 3) (6, 5)

Como pode ser observado neste caso as amostras (a, b) e (b, a) são consideradas diferentes, isto é, na amostragem com reposição as amostras são ordenadas.

**(2)** n = 3

Como N = 4 e n = 3, então o número de amostras possíveis será $N^n = 4^3 = 64$

Algumas destas amostras são:

(1, 1, 1) (1, 1, 3), (1, 3, 1), (3, 1, 1) (1, 3, 5) (1, 5, 3) ( 5, 3, 1) (5, 1, 3)

(1, 3, 6) (3, 3, 3), (5, 5, 5) (5, 5, 6) (1, 5, 6) (3, 5, 6) , etc.

## Estimador, estimativas e parâmetros

**Uma característica da população é denominada parâmetro.**

Um parâmetro é uma constante, isto é, é um número que representa uma característica única da população.

Assim se P é uma população, os principais parâmetros seriam:

(i) A média de P, anotada por $\mu_P$

(ii) A variância de P, anotada por $\sigma_P^2$

(iii) O desvio padrão de P, anotado por $\sigma_P$

(iv) A proporção de elementos de P que apresentam determinada característica, anotada por: $\pi$, entre outros.

Exemplo:

Para a população P = { 1, 3, 5, 6 } os parâmetros acima seriam:

(i) $\mu_P = (1 + 3 + 5 + 6) / 4 = 15 / 4 = 3{,}75$

(ii) $\sigma_P^2 = (1 + 9 + 25 + 36) / 4 - 3{,}75^2 = 71/4 - 3{,}75^2 = 17{,}75 - 14{,}0625 = 3{,}6875 = 3{,}69$.

(iv) $\sigma_P = 1{,}9203 = 1{,}92$

(v) $\pi = 1 / 4 = 25\%$, onde o numerador representa o número de elementos pares na população

## Estimador

**Um estimador é uma característica da amostra.**

Como a amostra é aleatória um estimador é uma variável aleatória. Assim tudo o que foi visto em probabilidade sobre variáveis aleatórias, aplica-se aos estimadores. A distribuição de probabilidade de um estimador é denominada de **distribuição amostral**.

Os principais estimadores são:

(I) A média da amostra, $\overline{X}$ que é um estimador da média da população: $\mu$

(ii) A variância amostral, $S^2$ que é um estimador da variância populacional: $\sigma^2$

(iii) A proporção amostral, P, que é um estimador amostral da proporção populacional $\pi$.

## Estimativa

**Uma estimativa é um valor particular de um estimador**

Assim $\overline{x} = 2$ é uma estimativa. O estimador é a expressão (fórmula) enquanto que a estimativa é o valor particular que ele assume (número).

## Cálculo dos principais estimadores.

Se $(X_1, X_2, ..., X_n)$ é uma amostra aleatória de tamanho "n" extraída de uma população, então:

**(a)** $\overline{X} = \sum X_i / n$ é uma estimativa da média populacional quando a amostra não está agrupada e $\overline{X} = \sum f_i X_i / n$ é uma estimativa da média da amostra quando a amostra está agrupada em uma distribuição de freqüências (por ponto ou por valores).

**(b)** $S^2 = \sum \frac{(X_i - \overline{X})^2}{n-1} = \sum \frac{X_i^2 - n\overline{X}^2}{n-1}$ é uma estimativa da variância populacional quando a amostra não está agrupada e

$S^2 = \sum \frac{f_i(X_i - \overline{X})^2}{n-1} = \sum \frac{f_i X_i^2 - n\overline{X}^2}{n-1}$ é uma estimativa da variância populacional quando a amostra está agrupada em uma distribuição de freqüências. Note-se que agora a variância é calculada com
"n - 1" no denominador. Isto se deve ao fato de que a variância for calculada com "n" no denominador, a média de sua distribuição amostral não será igual a variância populacional o que caracterizaria um estimador tendencioso.

Embora a variância seja calculada com "n - 1" no denominador com o objetivo de que as estimativas variem em torno do parâmetro, isto não irá ocorrer se a amostragem for sem reposição de população finita. Neste caso é necessário utilizar, ainda, uma correção para a variância que consiste em multiplicá-la pelo valor (N - 1) / N. Evidentemente esta correção só será necessária se a população for pequena, caso contrário o quociente acima será aproximadamente igual a um e a correção não precisará ser feita.

Assim se a população for finita (e pequena) e a amostragem for realizada **sem** reposição a variância deverá ser calculada por:

$$\hat{S}^2 = \frac{N-1}{N} S^2$$

**(c)** P = f / n, onde f = freqüência de elementos na amostra com determinada característica é uma estimativa da proporção populacional $\pi$.

## 1.2. DISTRIBUIÇÃO AMOSTRAL DOS ESTIMADORES

### 1.2.1. Distribuição amostral da média

#### (1) Amostragem com reposição

Considere-se a população P = { 1, 3, 5, 6 } e todas as amostras possíveis de tamanho n = 2 extraídas **com** reposição. Para cada amostra vai-se calcular a média. Ter-se-á assim um conjunto de 16 valores que serão dispostos em uma tabela, com as respectivas probabilidades, e que constituirá então a distribuição amostral da média da amostra.

As possíveis amostras com as respectivas médias são:

| Amostras | (1, 1) | (1, 3) | (1, 5) | (1, 6) | (3, 3) | (3, 5) | (3, 6) | (5, 5) |
|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **3,5** | **3** | **4** | **4,5** | **5** |
| Amostras | (5, 6) | (6, 6) | (3, 1) | (5, 1) | (6, 1) | ( 5, 3) | (6, 3) | (6, 5) |
| $\bar{x}$ | **5,5** | **6** | **2** | **3** | **3,5** | **4** | **4,5** | **5,5** |

Colocando estes resultados em uma tabela (distribuição amostral da média) vem:

| $\bar{x}$ | $f(\bar{x}) = P(\bar{X} = \bar{x})$ | $\bar{x}\,f(\bar{x})$ | $\bar{x}^2 f(\bar{x})$ |
|---|---|---|---|
| 1,0 | 1/16 | 1/16 | 1,0/16 |
| 2,0 | 2/16 | 4/16 | 8,0/16 |
| 3,0 | 3/16 | 9/16 | 27,0/16 |
| 3,5 | 2/16 | 7/16 | 24,5/16 |
| 4,0 | 2/16 | 8/16 | 32,0/16 |
| 4,5 | 2/16 | 9/16 | 40,5/16 |
| 5,0 | 1/16 | 5/16 | 25,0/16 |
| 5,5 | 2/16 | 11/16 | 60,5/16 |
| 6,0 | 1/16 | 6/16 | 36,0/16 |
| $\Sigma$ | **1** | **60/16** | **254,5/16** |

Pela tabela pode-se verificar que:

$E(\bar{X}) = \sum \bar{x} f(\bar{x}) = 60/16 = 3,75 = \mu$, isto é a expectância (média) de todas as médias amostrais, extraídas com reposição da população P, é igual a média populacional (parâmetro populacional média).

$V(\bar{X}) = \sum \bar{x}^2 f(\bar{x}) - \mu_{\bar{X}}^2 = 254,5/16 - 3,75^2 = 1,84375 = \sigma^2/\ 2 = 3,6875/2$, isto é, a variância entre as médias amostrais é "n" vezes (neste caso 2 vezes) menor que a variância populacional.

O valor $\sigma_{\bar{X}} = 1,36$ é denominado **erro padrão da média**. Ele mede a variabilidade entre as médias amostrais e dá uma idéia do erro que se comete ao se substituir a média da população pela média da amostra.

De fato, verificando a tabela acima, pode-se ver que se por exemplo, fosse selecionada a amostra (1, 1) seríamos levados a crer que a média da população seria um, quando de fato ela vale 3,75, cometendo assim um erro de 2,75 unidades. Felizmente este erro (o maior possível neste caso) só vai ocorrer com uma probabilidade de 1/16 = 6,25%. Se por exemplo, fosse selecionada a amostra (1,

6) a média amostral seria 3,5 e o erro cometido (neste caso) seria de 0,25 unidades. Este erro bem menor que o anterior ocorre com uma probabilidade de 2/16 = 12,5%. O que o desvio padrão da distribuição amostral da média faz é determinar o erro médio, sendo por isso denominado, então, de erro padrão da amostragem.

## (2) Amostragem sem reposição

Considere-se a população P = { 1, 3, 5, 6 } e todas as amostras possíveis de tamanho n = 2 extraídas sem reposição.

As possíveis amostras com as respectivas médias são:

| Amostras | (1, 3) | (1, 5) | (1, 6) | (3, 5) | (3, 6) | (5, 6) |
|---|---|---|---|---|---|---|
| $\bar{x}$ | **2** | **3** | **3,5** | **4** | **4,5** | **5,5** |

Colocando estes resultados em uma tabela (distribuição amostral da média) vem:

| $\bar{x}$ | $f(\bar{x}) = P(\bar{X} = \bar{x})$ | $\bar{x}f(\bar{x})$ | $\bar{x}^2 f(\bar{x})$ |
|---|---|---|---|
| 2,0 | 1 / 6 | 2,0 / 6 | 04,00 / 6 |
| 3,0 | 1 / 6 | 3,0 / 6 | 09,00 / 6 |
| 3,5 | 1 / 6 | 3,5 / 6 | 12,25 / 6 |
| 4,0 | 1 / 6 | 4,0 / 6 | 16,00 / 6 |
| 4,5 | 1 / 6 | 4,5 / 6 | 20,25 / 6 |
| 5,5 | 1 / 6 | 5,5 / 6 | 30,25 / 6 |
| $\Sigma$ | **1** | **22,5 / 6** | **91,75 / 6** |

Da tabela segue:

$E(\bar{X}) = \sum \bar{x} f(\bar{x})$ = 22,5/6 = 3,75 = $\mu$, isto é a expectância (média) de todas as médias amostrais, extraídas sem reposição da população P, também é igual a média populacional (parâmetro populacional média).

$V(\bar{X}) = \sum \bar{x}^2 f(\bar{x}) - \mu_{\bar{X}}^2$ = 91,75/6 - 3,75$^2$ = 1,2292 = $\frac{\sigma^2}{2}.\frac{N-n}{N-1}$ = 1,84375. (2/3), isto é, a variância entre as médias amostrais é “n” vezes (neste caso 2 vezes) menor que a variância populacional multiplicada pelo fator de correção de população finita. Este fator, pode ser considerado como o fator de eficiência da amostragem sem reposição sobre a amostragem com reposição, que neste caso (N = 4 e n =2) vale 2/3. Como na amostragem sem reposição não é possível retirar o mesmo elemento duas vezes, as médias não podem assumir valores tão extremos, como por exemplo, o valor “um” ou “seis” que assumiram na amostragem com reposição. Isto faz com que a erro padrão na amostragem **sem** reposição seja menor do que na amostragem **com** reposição.

O fator de redução da variância na amostragem sem reposição é: (N - n) / (N -1)

Pode-se perceber facilmente que quanto maior for a diferença entre o tamanho da população e o tamanho da amostra mais próximo de “um” será este fator. Então, como regra prática, pode-se admitir como necessária a correção para a variância das médias amostrais sempre que o tamanho da amostra exceder a 5% do tamanho da população. Caso isto não ocorra não é necessário fazer-se a distinção entre os dois procedimentos (com e sem reposição).

Evidentemente as considerações acima valem para populações pequenas. Se a população é bastante grande ou infinita, não mais será possível pensar em construir tabelas para representar a distri-

buição das médias amostrais. Neste caso é necessário procurar por modelos probabilísticos que descrevam a distribuição da média amostral. Neste caso, também, como declarado acima a distinção entre amostragem com e sem reposição não será necessário, pois o fator de correção será "aproximadamente um" e não necessitará ser utilizado.

Os modelos probabilísticos são conhecidos a partir dos dois seguintes resultados:

(a) Se $(X_1, X_2, ..., X_n)$ é uma amostra aleatória de uma população com distribuição normal de média $\mu$ e desvio padrão $\sigma$, então a média da amostra ($\overline{X}$) terá uma distribuição também normal com a mesma média da população e com desvio padrão (erro padrão) raiz de "n" vezes menor que o desvio padrão da população, isto é:

Se X é $N(\mu, \sigma)$ então $\overline{X}$ será $N(\mu, \sigma/\sqrt{n})$

(b) Teorema Central do Limite

Se $(X_1, X_2, ..., X_n)$ é uma amostra aleatória extraída de uma população com qualquer distribuição de média $\mu$ e desvio padrão $\sigma$, então a média da amostra ($\overline{X}$) terá uma distribuição aproximadamente normal com a mesma média da população e com desvio padrão (erro padrão) raiz de "n" vezes menor que o desvio padrão da população à medida que o tamanho da amostra aumenta.

OBS.: Para amostras de 30 ou mais valores, em geral, a aproximação já será suficiente boa, para se poder utilizar este resultado.

Assim:

Se X tem **qualquer** distribuição então $\overline{X}$ terá uma distribuição aproximadamente $N(\mu, \sigma/\sqrt{n})$ para n grande ($n \geq 30$).

## Exemplos:

(1) Uma população X tem uma distribuição normal de média 100 e desvio padrão 10.

(a) Qual $P(95 < X < 105)$?

(b) Se $\overline{X}$ é a média de 16 elementos extraída desta população, qual a $P(95 < \overline{X} < 105)$ ?

## Solução:

(a) Como X é uma N(100, 10) vem:

$P(95 < X < 105) = P(-0,5 < Z < 0,5) = \Phi(0,5) - \Phi(-0,5) = 0,6915 - 0,3185 = 38,30\%$.

Neste caso $\overline{X}$ é uma N(100; 2,5), então:

(b) $P(95 < \overline{X} < 105) = P(-2,0 < Z < 2,0) = \Phi(2,0) - \Phi(-2,0) = 0,9772 - 0,0228 = 95,44\%$.

(2) A renda de um conjunto de pessoas de uma certa região tem média 6 s.m. e desvio padrão de 2 s.m. Se desta população for extraída uma amostra de n = 100 pessoas, qual a probabilidade de a média desta amostra acuse um valor superior a 6,3 s.m?

## Solução:

Neste caso, como não foi declarado que a população é normal é necessário aplicar o teorema central do limite, uma vez que n = 100 > 30, isto é possível. A média da amostra terá uma distribuição aproximadamente normal com média 6 s.m. e desvio padrão de: 2 / 10 = 0,20, uma vez que o erro pa-

drão da média é raiz de n vezes menor do que o desvio padrão populacional. Então, a probabilidade pedida será:

$P(\overline{X} > 6{,}30) = P(Z > (6{,}30 - 6)/0{,}20\ ) = P\ (Z > 1{,}5) = \Phi(-1{,}5) = 6{,}68\%$, isto é, apenas 6,68% das médias de amostras de tamanho n = 100 apresentarão um valor superior a 6,30 s.m.

## 1.2.2. Distribuição amostral da variância

### (1) Amostragem COM reposição

Considere-se a população P = { 1, 3, 5, 6 } e todas as amostras possíveis de tamanho n = 2 extraídas com reposição. Para cada amostra vai-se calcular a variância. Ter-se-á assim um conjunto de 16 valores que serão dispostos em uma tabela, com as respectivas probabilidades, e que constituirá então a distribuição amostral da variância.

As possíveis amostras com as respectivas variâncias são:

| Amostras | (1, 1) | (1, 3) | (1, 5) | (1, 6) | (3, 3) | (3, 5) | (3, 6) | (5, 5) |
|---|---|---|---|---|---|---|---|---|
| $\overline{x}$ | 1 | 2 | 3 | 3,5 | 3 | 4 | 4,5 | 5 |
| $s^2$ | **0** | **2** | **8** | **12,5** | **0** | **2** | **4,5** | **0** |
| Amostras | (5, 6) | (6, 6) | (3, 1) | (5, 1) | (6, 1) | ( 5, 3) | (6, 3) | (6, 5) |
| $\overline{x}$ | 5,5 | 6 | 2 | 3 | 3,5 | 4 | 4,5 | 5,5 |
| $s^2$ | **0,5** | **0** | **2** | **8** | **12,5** | **2** | **4,5** | **0,5** |

Colocando estes resultados em uma tabela(distribuição amostral da variância) vem:

| $s^2$ | $f(s^2) = P(S^2 = s^2)$ | $s^2f(s^2)$ |
|---|---|---|
| 0,0 | 4/16 | 0/16 |
| 0,5 | 2/16 | 1/16 |
| 2,0 | 4/16 | 8/16 |
| 4,5 | 2/16 | 9/16 |
| 8,0 | 2/16 | 16/16 |
| 12,5 | 2/16 | 25/16 |
| $\sum$ | **1** | **59/16** |

Pela tabela segue que:

$E(S^2) = \sum s^2 f(s^2) = 59/16 = 3{,}6875 = \sigma^2$, isto é a expectância (média) de todas as variâncias das amostras de tamanho n = 2, extraídas **com** reposição da população P, é igual a variância populacional (parâmetro populacional variância). Em outras palavras, pode-se dizer que quando a amostragem é com reposição a variância amostral $S^2$ é um estimador não tendencioso da variância populacional $\sigma^2$.

Desta forma, sempre que se desejar estimar a variância de uma população onde as amostras foram retiradas **com** reposição, pode-se usar a variância amostral como estimador.

### (2) Amostragem SEM reposição

Considere-se a população P = { 1, 3, 5, 6 } e todas as amostras possíveis de tamanho n = 2 obtidas sem reposição.

As possíveis amostras com as respectivas variâncias são:

| Amostras | (1, 3) | (1, 5) | (1, 6) | (3, 5) | (3, 6) | (5, 6) |
|---|---|---|---|---|---|---|
| $\bar{x}$ | 2 | 3 | 3,5 | 4 | 4,5 | 5,5 |
| $s^2$ | **2** | **8** | **12,5** | **2** | **4,5** | **0,5** |

Colocando estes resultados em uma tabela (distribuição amostral da variância) vem:

| $s^2$ | $f(s^2) = P(S^2 = s^2)$ | $s^2f(s^2)$ |
|---|---|---|
| 0,5 | 1/6 | 0,5/6 |
| 2,0 | 2/6 | 4,0/6 |
| 4,5 | 1/6 | 4,5/6 |
| 8,0 | 1/6 | 8,0/6 |
| 12,5 | 1/6 | 12,5/6 |
| $\sum$ | **1** | **29,5/6** |

Pela tabela pode-se ver que:

$E(S^2) = \sum s^2 f(s^2) = 29{,}5/6 \neq 3{,}6875 = \sigma^2$, isto é a expectância (média) de todas as variâncias das amostras de tamanho n = 2, extraídas **sem** reposição da população finita P, não é igual a variância populacional (parâmetro populacional variância). Neste caso, para que se obtenha um estimador não tendencioso da variância populacional é necessário corrigir a variância amostral através do fator (N - 1) / N. Assim se cada variância acima for multiplicada por este fator, que neste caso será, (N - 1) / N = 3 / 4 = 0,75, então, se terá:

| $\hat{s}^2$ | $f(\hat{s}^2) = P(\hat{S}^2 = \hat{s}^2)$ | $\hat{s}^2 f(\hat{s}^2)$ |
|---|---|---|
| 0,375 | 1/6 | 0,375/6 |
| 1,500 | 2/6 | 3,000/6 |
| 3,375 | 1/6 | 3,375/6 |
| 6,000 | 1/6 | 6,000/6 |
| 9,375 | 1/6 | 9,375/6 |
| $\sum$ | **1** | **22,125/6** |

$E(\hat{S}^2) = \sum \hat{s}^2 f(\hat{s}^2) = 22{,}125 / 6 = 3{,}6875 = \sigma^2$, isto é a expectância (média) de todas as variâncias corrigidas é igual ao parâmetro populacional $\sigma^2$. Assim quando a população é pequena e amostragem for **sem** reposição é necessário corrigir a variância da amostra pelo fator **(N - 1) / N**, para que ela seja um bom estimador da variância populacional. É claro que esta correção só será importante para populações pequenas. Se a população for grande, por exemplo, N = 1000, então o fator (N - 1) / N = 999 / 1000 = 0,999 o que é aproximadamente 1. Neste caso, não é necessário usar esta correção e a amostragem sem reposição pode ser considerada equivalente a com reposição para efeitos de estimação da variância populacional.

Evidentemente as considerações acima valem para populações pequenas. Se a população é bastante grande ou infinita, não mais será possível pensar em construir tabelas para representar a distribuição das variâncias amostrais. Neste caso é necessário procurar por modelos probabilísticos (funções) que descrevam a distribuição da variância amostral. Para a variância este modelo existe e é denominado de distribuição Qui-quadrado ($\chi^2$).

## 1.2.3. Distribuição amostral da proporção

### (1) Amostragem COM reposição

Considere-se a população P = { 1, 3, 5, 6 } e todas as amostras possíveis de tamanho n = 2 obtidas **com** reposição. Para cada amostra vai-se calcular a proporção P de elementos pares na população. Ter-se-á assim um conjunto de 16 valores que serão dispostos em uma tabela, com as respectivas probabilidades, e que formarão então a distribuição amostral da proporção.

As possíveis amostras com as respectivas proporções são:

| Amostras | (1, 1) | (1, 3) | (1, 5) | (1, 6) | (3, 3) | (3, 5) | (3, 6) | (5, 5) |
|---|---|---|---|---|---|---|---|---|
| **p** | **0** | **0** | **0** | **1/2** | **0** | **0** | **1/2** | **0** |
| Amostras | (5, 6) | (6, 6) | (3, 1) | (5, 1) | (6, 1) | ( 5, 3) | (6, 3) | (6, 5) |
| **p** | **1/2** | **1** | **0** | **0** | **1/2** | **0** | **1/2** | **1/2** |

Colocando estes resultados em uma tabela (distribuição amostral da proporção) vem:

| **p** | **f(p) = P(P = p)** | **pf(p)** | **$p^2f(p)$** |
|---|---|---|---|
| 0,0 | 9/16 | 0/16 | 0,0/16 |
| 0,5 | 6/16 | 3/16 | 1,5/16 |
| 1,0 | 1/16 | 1/16 | 1,0/16 |
| **Σ** | **1** | **4/16** | **2,5/16** |

Pode-se então calcular a expectância e a variância:

$E(P) = \sum pf(p) = 4/16 = 0,25 = \pi$, isto é o valor esperado (média) de todas as proporções amostrais, extraídas **com** reposição da população P, e é igual a proporção populacional (parâmetro populacional $\pi$). Isto significa, que o estimador P é um estimador não tendencioso (ou não viciado) da proporção populacional $\pi$, quando as amostras são extraídas **com** reposição da população.

$V(P) = \sum p^2f(p) - \mu_P^2 = 2,5/16 - 0,25^2 = 0,09375 = \pi(1 - \pi) / n$, isto é, a variância entre as proporções amostrais é "n" vezes (neste caso 2 vezes) menor que a variância populacional. Isto porque quando se está trabalhando com proporções, pode-se mostrar que a variância populacional é igual a $\pi(1 - \pi)$.

O valor $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} = 0,09375$ é denominado **erro padrão da proporção**. Ele mede a variabilidade entre as proporções amostrais e dá uma idéia do erro que se comete ao se substituir a proporção da população pela proporção da amostra.

### (2) Amostragem SEM reposição

Considere-se a população P = { 1, 3, 5, 6 } e todas as amostras possíveis de tamanho n = 2 extraídas **sem** reposição.

As possíveis amostras com as respectivas proporções são:

| Amostras | (1, 3) | (1, 5) | (1, 6) | (3, 5) | (3, 6) | (5, 6) |
|---|---|---|---|---|---|---|
| **p** | **0** | **0** | **1/2** | **0** | **1/2** | **1/2** |

Colocando estes resultados em uma tabela (distribuição amostral da proporção) vem:

| p | f(p) = P(P = p) | pf(p) | $p^2f(p)$ |
|---|---|---|---|
| 0,0 | 1/2 | 0,0/2 | 0,00/2 |
| 0,5 | 1/2 | 0,5/2 | 0,25/2 |
| $\sum$ | **1** | **0,5/2** | **0,25/2** |

Portanto:

$E(P) = \sum pf(p) = 0,5/2 = 0,25 = \pi$, isto é a expectância (média) de todas as proporções amostrais, extraídas **sem** reposição da população P, e é igual a proporção populacional (parâmetro populacional $\pi$). Isto significa, que o estimador P é um estimador não tendencioso (ou não viciado) da proporção populacional $\pi$, quando as amostras são retiradas **sem** reposição.

$V(P) = \sum p^2f(p) - \mu_P^2 = 0,25/2 - 0,25^2 = 0,0625 = \frac{\pi(1-\pi)}{2}.\frac{N-n}{N-1}$, isto é, a variância entre as proporções amostrais é “n” vezes (neste caso 2 vezes) menor que a variância populacional multiplicada pelo fator de correção de população finita. Este fator, pode ser considerado como o fator de eficiência da amostragem sem reposição sobre a amostragem com reposição que, neste exemplo, (N = 4 e n = 2) vale 2/3.

Evidentemente as considerações acima valem para populações pequenas. Se a população é bastante grande ou infinita, não mais será possível pensar em construir tabelas para representar a distribuição das proporções amostrais. Nesta situação é necessário procurar por modelos probabilísticos que descrevam a distribuição da proporção amostral. Neste caso, também, como declarado acima a distinção entre amostragem com e sem reposição não será necessária, pois o fator de correção será “aproximadamente um” e não precisará ser utilizado.

O modelo probabilístico para a proporção amostral é dada pelo seguinte resultado:

(a) Se $(X_1, X_2, ..., X_n)$ é uma amostra aleatória retirada de uma população com proporção $\pi$, então a distribuição da proporção amostral será aproximadamente normal com média $\mu_P = \pi$ e desvio padrão $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$.

OBS.: Para amostras de 30 ou mais valores, em geral, a aproximação já será suficiente boa, para se poder utilizar este resultado. Para amostras pequenas a distribuição da proporção amostral é Binomial.

## Exemplo:

(1) A proporção de eleitores do candidato D. M. A. Gogo numa certa região é de 20%. Extraída uma amostra de 100 eleitores desta região, qual a probabilidade que ela apresente um número de eleitores do candidato

(a) Abaixo de 15%	(b) Superior a 30%

**Solução:**

Como n > 30 pode-se usar a distribuição normal com média $\mu = \pi = 20\%$ e desvio padrão $\sigma = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0,2(1-0,2)}{100}} = 0,04 = 4\%$, Então:

(a) $P(P < 15\%) = P(Z < -1,25) = \Phi(-1,25) = 10,56\%$.

(b) $P(P > 30) = P(Z > 2,5) = \Phi(-2,5) = 0,62\%$.

# 2. ESTIMAÇÃO

A inferência estatística tem por objetivo fazer generalizações sobre uma população com base em valores amostrais. A inferência pode ser feita estimando os parâmetros:

(a) Por ponto e

(b) Por intervalo.

A estimação por ponto é feita através de um único valor, enquanto que a estimação por intervalo fornece um intervalo de valores em torno do valor da estimativa pontual.

<u>Exemplo:</u>

Uma amostra aleatória simples de 400 pessoas de uma cidade é extraída e 300 respondem que acham a administração municipal boa ou ótima. Então o valor p = 300/400 = 75% é uma estimativa por ponto do percentual de pessoas da cidade que acham a administração boa ou ótima. Esta mesma estimativa poderia ser enunciado como de: 70% a 80% das pessoas da cidade acham a administração boa ou ótima. Neste caso, teríamos uma estimativa por intervalo da proporção. Note-se que o centro do intervalo é o valor "75%" da estimativa pontual.

## 2.1. PROPRIEDADES DOS ESTIMADORES

Seja X uma população com um parâmetro de interesse $\theta$ e seja ($X_1$, $X_2$, ..., $X_n$) uma amostra aleatória simples extraída desta população. Seja $\hat{\theta}$ um estimador do parâmetro $\theta$. Então:

(i) Se $E(\hat{\theta}) = \theta$ se dirá que $\hat{\theta}$ é um estimador não-tendencioso ou não viciado do parâmetro populacional $\theta$. Neste caso, a média do estimador $\hat{\theta}$ é o parâmetro populacional $\theta$, ou ainda, pode-se dizer que o estimador varia em torno do parâmetro populacional.

(ii) Se $\hat{\theta}$ é um estimador não tendencioso de um parâmetro $\theta$, se dirá que $\hat{\theta}$ é consistente se à medida que o tamanho da amostra aumenta a variabilidade do estimador diminui, isto é, as observações vão ficando cada vez mais concentradas em torno do parâmetro na medida em que a amostra vai ficando cada vez maior. Em símbolos:

$$\lim_{n \to \infty} V(\hat{\theta}) = 0$$

## 2.2. ESTIMAÇÃO POR PONTO

Seja X uma população com média $\mu$, desvio padrão $\sigma$ e com uma proporção $\pi$ e seja ($X_1$, $X_2$, ..., $X_n$) uma amostra aleatória simples extraída desta população. Então:

(a) $\overline{X}$ é um estimador não-tendencioso e consistente da média da população $\mu$.

(b) P é um estimador não-tendencioso e consistente da proporção populacional $\pi$.

(c) $S^2$ é estimador não-tendencioso e consistente da variância da população $\sigma^2$, a menos que a extração seja **sem** reposição de população finita. Neste caso, o estimador é $\hat{S}^2 = \frac{N-1}{N} S^2$.

## 2.3. ESTIMAÇÃO POR INTERVALO

O estimador por ponto não permite ter uma idéia do erro cometido ao se fazer a estimativa do parâmetro. Para que se possa associar uma confiança (probabilidade) a uma estimativa é necessário

construir um intervalo em torno da estimativa por ponto. Este intervalo é construído baseado na distribuição amostral do estimador.

## 2.3.1. Da média populacional

### (a) Desvio padrão populacional (σ) conhecido

O intervalo de confiança para a média (μ) de uma população é construído em torno da estimativa pontual $\overline{X}$. Para construir este intervalo fixa-se uma probabilidade "1 - α" de que o intervalo construído contenha o parâmetro populacional. Desta forma, "α" será a probabilidade de que o intervalo obtido não contenha o valor do parâmetro, isto é, "α" será a probabilidade de erro. Sabe-se que a média da amostra tem distribuição normal de média μ e desvio padrão $\sigma/\sqrt{n}$ se a população de onde for extraída a amostra for normal (ou se a amostra for superior a 30 e retirada de qualquer população ) de média μ e de desvio padrão σ, pode-se então utilizar a curva normal para estabelecer os limites para o intervalo de confiança.

Lembrando que o que se quer é um intervalo que contenha o parâmetro populacional μ com probabilidade "1 - α" tem-se então:

$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$, onde $z_{\alpha/2}$ é o valor da normal padrão com área à direita é igual a α/2.

Mas $Z = (\overline{X} - \mu) / \sigma/\sqrt{n}$ substituindo na expressão acima vem:

$P(-z_{\alpha/2} < (\overline{X} - \mu) / \sigma/\sqrt{n} < z_{\alpha/2}) = 1 - \alpha$. Trabalhando esta desigualdade, segue que:

$P(\overline{X} - z_{\alpha/2}\,\sigma/\sqrt{n} < \mu < \overline{X} + z_{\alpha/2}\,\sigma/\sqrt{n}) = 1 - \alpha$. Que é o intervalo procurado. Assim o intervalo de confiança (probabilidade) de "1 - α" para a média de uma população é dado por:

$[\overline{X} - z_{\alpha/2}\,\sigma/\sqrt{n} ; \overline{X} + z_{\alpha/2}\,\sigma/\sqrt{n}]$ onde:

$\overline{X}$ é a estimativa por ponto da média da população.

σ é o desvio padrão da população e

$z_{\alpha/2}$ é o valor da distribuição normal padrão cuja área à direita é igual a α/2, isto é, é o valor de Z tal que: $P(Z > z_{\alpha/2}) = \alpha/2$, ou então: $\Phi(-z_{\alpha/2}) = \alpha/2$.

Exemplo:

Uma população tem um desvio padrão igual a 10 e média desconhecida. Uma amostra de tamanho n = 100 é retirada e fornece uma média $\overline{x}$ = 50. Qual o intervalo de 95% de confiança para a média desta população?

Solução:

Tem-se 1 - α = 95%, então α = 5% e α / 2 = 2,5%. O coeficiente de confiança que deve ser buscado na normal padrão é valor $z_{\alpha/2}$ de Z tal que:

$P(Z > z_{\alpha/2}) = 2,5\%$, ou então: $\Phi(-z_{\alpha/2}) = 2,5\%$.

Este valor vale 1,96. Então o intervalo de confiança de 95% para a média desta população será:

$[\overline{X} - z_{\alpha/2} \sigma/\sqrt{n} ; \overline{X} + z_{\alpha/2} \sigma/\sqrt{n}]$ = [50 - 1,96.10/10; 50 + 1,96.10/10] = [50 - 1,96; 50 + 1,96] =

[48,04; 51,96], ou seja, pode-se afirmar com uma certeza de 95% de que este intervalo conterá a média desta população.

Obs.: O valor $\varepsilon = z_{\alpha/2} \sigma/\sqrt{n}$ é denominado de erro padrão da estimação. Não confundir com o valor $\sigma/\sqrt{n}$ que é o erro padrão da amostragem. O erro padrão da estimação é a semi-amplitude do intervalo de confiança. A amplitude do intervalo de confiança (IC) será; $2\varepsilon$.

### (b) Desvio padrão populacional ($\sigma$) desconhecido

Quando o desvio padrão da população ($\sigma$) é desconhecido é necessário utilizar sua estimativa "s". Só que ao substituir-se o desvio padrão populacional pelo sua estimativa no quociente:

$(\overline{X} - \mu) / \sigma/\sqrt{n}$ não se terá mais uma normal padrão. De fato, conforme demonstrado pelo estatístico inglês W. S. Gosset, conhecido por "Student" o comportamento do quociente:

$(\overline{X} - \mu) / S/\sqrt{n}$ segue uma distribuição simétrica em torno de zero, porém com uma variabilidade maior do que a da normal padrão. A distribuição do quociente acima é conhecida como distribuição "t" de Student.

Na realidade existem infinitas distribuições "t", uma para cada tamanho de amostra. Estas distribuições a exemplo da normal padrão encontram-se tabeladas.

A tabela para a distribuição "t" segue uma metodologia um pouco diferente daquela da normal padrão. De fato, como existem muitas distribuições de Student não seria possível tabelá-las da mesma forma que a da normal padrão. Assim cada linha de uma tabela representa uma distribuição diferente e cada coluna representa um valor de confiança que poderá ser "$\alpha$" ou "$\alpha/2$", isto é, a tabela poderá ser unilateral ou bilateral. A linha de cada tabela fornece a distribuição "t" com parâmetro "n - 1" denominado de graus de liberdade, isto é, o grau de liberdade = $\nu$ = n - 1 = linha da tabela.

Neste caso, o intervalo de confiança com probabilidade "1 - $\alpha$" para a média será:

$[\overline{X} - t_{\alpha/2} S/\sqrt{n} ; \overline{X} + t_{\alpha/2} S/\sqrt{n}]$ onde:

$\overline{X}$ é a estimativa por ponto da média da população;

S é o desvio padrão da amostra e uma estimativa do desvio padrão da população $\sigma$ e

$t_{\alpha/2}$ é o valor da distribuição t cuja área à direita é igual a $\alpha/2$, isto é, é o valor de t tal que:

$P(t > t_{\alpha/2}) = \alpha/2$, ou então: $P(- t_{\alpha/2} < t < t_{\alpha/2}) = 1 - \alpha$.

## Exemplo:

Uma amostra de tamanho 25 foi retirada de uma população com o objetivo de estimar a sua média e forneceu os valores $\overline{x}$ = 50 e s = 10. Qual o intervalo de 95% de confiança para a média desta população?

Solução:

Tem-se $1 - \alpha = 95\%$, então $\alpha = 5\%$ e $\alpha / 2 = 2,5\%$. O coeficiente de confiança que deve ser buscado na distribuição t com $\nu = n - 1 = 25 - 1 = 24$. Esta é a linha da tabela. A coluna poderá ser o valor $\alpha = 5\%$ ou então o valor $\alpha / 2 = 2,5\%$, dependendo do tipo de tabela. Em qualquer caso o que se procura é o valor "t" com grau de liberdade igual a 24, isto é, o valor $t_{24}$ tal que:

$P(- t_{\alpha/2} < t_{24} < t_{\alpha/2}) = 95\%$

Este valor vale 2,064. (Note-se que na a normal este mesmo valor valia 1,96). Então o intervalo de confiança de 95% para a média desta população será:

$[\overline{X} - t_{\alpha/2} S/\sqrt{n} ; \overline{X} + t_{\alpha/2} S/\sqrt{n}] = [50 - 2,064.10/5; 50 + 2,064.10/5] = [50 - 4,13; 50 + 4,13] = [45,87; 54,13]$, ou seja, pode-se afirmar com uma certeza de 95% de que este intervalo conterá a média desta população.

Convém notar que a última linha da tabela da distribuição "t" apresenta valores coincidentes com aqueles que seriam obtidos se fosse utilizado a distribuição normal padrão. Isto ocorre porque a distribuição "t" tende a distribuição normal à medida que o tamanho da amostra aumenta, isto é, a distribuição normal é o limite da distribuição "t" quando o tamanho da amostra tende ao infinito.

Esta aproximação já será bastante boa para amostras de tamanho $n > 30$. Assim se a amostra for superior a 30 pode-se utilizar a distribuição normal ao invés da distribuição "t", isto é, pode-se ler os valores na normal padrão, ou então na última linha da tabela "t".

## 2.3.2. Da proporção populacional

Seja P = proporção amostral. Sabe-se que para $n > 30$ a distribuição amostral de P é aproximadamente normal com média $\mu_P = \pi$ e desvio padrão (erro padrão) $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$. Pode-se então utilizar a curva normal para estabelecer os limites para o intervalo de confiança.

Lembrando que o que se quer é um intervalo que contenha o parâmetro populacional $\pi$ com probabilidade "$1 - \alpha$" então tem-se:

$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$, onde $z_{\alpha/2}$ é o valor da normal padrão com área à direita é igual a $\alpha/2$.

Mas $Z = (P- \mu_P) / \sigma_P$ então substituindo na expressão acima vem:

$P(-z_{\alpha/2} < (P - \mu_P) / \sigma_P < z_{\alpha/2} ) = 1 - \alpha$. Trabalhando esta desigualdade, segue que:

$P(P - z_{\alpha/2}\sigma_P < \mu_P < P + z_{\alpha/2}\sigma_P) = P(P - z_{\alpha/2}\sigma_P < \pi < P + z_{\alpha/2}\sigma_P) = 1 - \alpha$. Que é o intervalo procurado. Assim o intervalo de confiança (probabilidade) de "$1 - \alpha$" para a proporção "P" de uma população é dado por:

$$[P- z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} ; P + z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}].$$

Observando-se a expressão acima pode-se perceber que o intervalo de confiança para a proporção populacional $\pi$, depende dele mesmo, isto é, é necessário calcular o erro amostral que está expresso em função de $\pi$. Como o objetivo é estimar este valor, evidentemente ele não é conhecido. As-

sim é necessário utilizar, sua estimativa $\hat{\sigma}_P$, isto é, é necessário substituir $\pi$ por P na expressão $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$. Desta forma o intervalo acima ficará:

$$[P - z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}\ ;\ P + z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}]$$, onde:

P é a estimativa por ponto da proporção populacional $\pi$.

$\hat{\sigma}_P = \sqrt{\frac{P(1-P)}{n}}$ é uma estimativa do erro padrão, isto é, do desvio padrão amostral e

$z_{\alpha/2}$ é o valor da distribuição normal padrão cuja área à direita é igual a $\alpha/2$. É o valor de Z tal que: $P(Z > z_{\alpha/2}) = \alpha/2$, ou então: $\Phi(-z_{\alpha/2}) = \alpha/2$.

## Exemplo 1:

Numa pesquisa de mercado, 400 pessoas foram entrevistadas sobre sua preferência por determinado produto. Destas 400 pessoas, 240 disseram preferir o produto. Determinar um intervalo de confiança de 95% de probabilidade para o percentual de preferência dos consumidores em geral para este produto.

## Solução:

Tem-se $1 - \alpha = 95\%$, então $\alpha = 5\%$ e $\alpha / 2 = 2{,}5\%$. O coeficiente de confiança que deve ser buscado na normal padrão é valor $z_{\alpha/2}$ de Z tal que:

$P(Z > z_{\alpha/2}) = 2{,}5\%$, ou então: $\Phi(-z_{\alpha/2}) = 2{,}5\%$.

Este valor vale 1,96. A estimativa por ponto para a proporção populacional será: $p = f/n = 240/400 = 0{,}60 = 60\%$.

Então o intervalo de confiança de 95% para a proporção populacional será:

$$[P - z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}\ ;\ P + z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}] = [0{,}60 - 1{,}96\sqrt{\frac{0{,}60(1-0{,}60)}{400}}\ .;\ 0{,}60 + 1{,}96\sqrt{\frac{0{,}60(1-0{,}60)}{400}}]$$

= [60% - 4,80% ; 60% + 4,80%] = [55,20%; 64,80%], ou seja, pode-se afirmar com uma certeza de 95% de que este intervalo conterá a proporção populacional, isto é, a verdadeira percentagem dos consumidores que preferem o produto pesquisado.

## Exemplo 2:

Numa pesquisa de mercado para estudar a preferência da população de uma cidade em relação ao consumo de um determinado produto, colheu-se uma amostra aleatória de 300 consumidores da cidade e observou-se que 180 consumiam o produto. Determinar um IC de 99% para a proporção populacional de consumidores do produto.

## Solução:

Tem-se $1 - \alpha = 99\%$, então $\alpha = 1\%$ e $\alpha / 2 = 0{,}5\%$. O coeficiente de confiança que deve ser buscado na normal padrão é valor $z_{\alpha/2}$ de Z tal que:

$P(Z > z_{\alpha/2}) = 0{,}5\%$, ou então: $\Phi(-z_{\alpha/2}) = 0{,}5\%$.

Este valor vale 2,575. A estimativa por ponto para a proporção populacional será: $p = f/n = 180/300 = 0{,}60 = 60\%$.

Então o intervalo de confiança de 99% para a proporção populacional será:

$[P- z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}$ ; $P + z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}] = [0,60 - 2,58\sqrt{\frac{0,60(1-0,60)}{300}}$ .; $0,60 + 2,58\sqrt{\frac{0,60(1-0,60)}{300}}]$ = [60% - 7,28% ; 60% + 7,28%] = [52,72%; 67,28%], ou seja, pode-se afirmar com uma certeza de 99% de que este intervalo conterá a proporção populacional, isto é, a verdadeira percentagem dos consumidores que preferem o produto pesquisado.

## 2.3.3. Da variância populacional (σ2)

Sabe-se que o estimador não-tendencioso de $\sigma^2$ é $S^2$ e que $E(S^2) = \sigma^2$, enquanto $V(S^2) = 2\sigma^2/(n-1)$. No entanto, para se construir um intervalo de confiança para $\sigma^2$ é necessário, ainda conhecer qual é o comportamento de $S^2$, isto é, qual é o modelo teórico (probabilístico) seguido pelo estimador. Assim antes de se construir um intervalo de confiança para a variância populacional é necessário se conhecer um novo modelo probabilístico denominado de qui-quadrado e representado por $\chi^2$ (c grego).

### A distribuição qui-quadrado

A distribuição ou modelo qui-quadrado pode ser obtida de uma soma de variáveis normais padronizadas, isto é, $\chi_n^2 = \sum_{i=1}^{n} Z_i^2$.

A distribuição $\chi^2$ é assimétrica positiva (possuí uma cauda à direita) e de depende do parâmetro ν. Sabe-se também que:

$E(\chi^2) = \nu$ e que $V(\chi^2) = 2\nu$.

A figura 2.1 mostra alguns exemplos de modelos qui-quadrado.

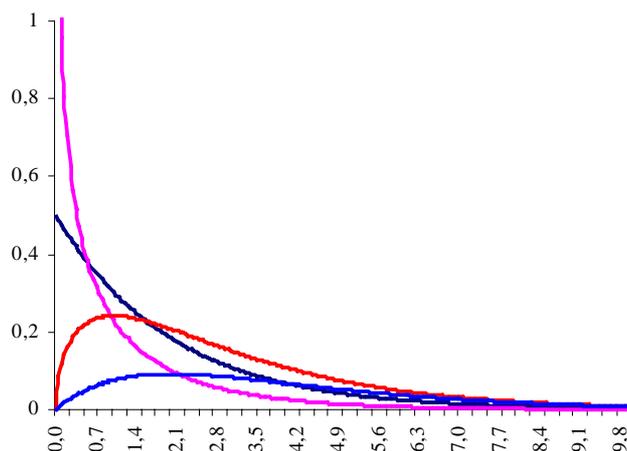



**Figura 2.1 - Algumas distribuições qui-quadrado**

A comportamento, distribuição de probabilidade, apresentado pela variância amostral ($S^2$) está relacionado com a distribuição (modelo) $\chi^2$ através do seguinte resultado:

$\chi^2_{n-1} = \frac{(n-1)S^2}{\sigma^2}$, isto é, a variância segue uma distribuição $\chi^2$ com "n - 1" graus de liberdade a menos de uma constante. Neste caso $\nu$ = n -1.

### Tabelas

A distribuição $\chi^2$ está tabelada em função do grau de liberdade n - 1 = $\nu$ (linha da tabela) e área à sua direita, isto é, $P(\chi^2 > c) = \alpha$. Na realidade o que está tabelado é a função inversa da $\chi^2$, isto é, entrando com o valor do parâmetro (graus de liberdade) e uma determinada probabilidade (área), a tabela fornece um valor da variável (abscissa) tal que a probabilidade à direita (área) deste valor seja igual a área especificada.

### O intervalo

Suponha que seja fixado um nível de confiança de “1 - $\alpha$“ e que $\chi^2_1$ e $\chi^2_2$ sejam dois valores da distribuição $\chi^2$ tais que $P(\chi^2_1 < \chi^2 < \chi^2_2) = 1 - \alpha$.

$$P(\chi^2_1 < \chi^2 < \chi^2_2) = 1 - \alpha$$

$$P\left(\chi^2_1 < \frac{(n-1)S^2}{\sigma^2} < \chi^2_2\right) = 1 - \alpha$$

$$P\left(\frac{1}{\chi^2_2} < \frac{\sigma^2}{(n-1)S^2} < \frac{1}{\chi^2_1}\right) = 1 - \alpha$$

$$P\left(\frac{(n-1)S^2}{\chi^2_2} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_1}\right) = 1 - \alpha$$

Assim o intervalo de confiança (probabilidade) de “1 - $\alpha$“ para a variância da população é dado por:

$$\left[\frac{(n-1)S^2}{\chi^2_2}; \frac{(n-1)S^2}{\chi^2_1}\right]$$

## 2.3.4. Do desvio padrão populacional ($\sigma$)

Para determinar um intervalo de confiança de "1 - $\alpha$" de probabilidade para o desvio padrão populacional basta apenas tomar a raiz quadrada positiva dos termos do intervalo para a variância populacional. Assim o intervalo será:

$$\left[\sqrt{\frac{(n-1)S^2}{\chi^2_2}}; \sqrt{\frac{(n-1)S^2}{\chi^2_1}}\right]$$

O significado deste intervalo é:

$$P\left(\sqrt{\frac{(n-1)S^2}{\chi_2^2}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi_1^2}}\right) = 1-\alpha\ .$$

## Exemplo:

Uma amostra extraída de uma população normal forneceu uma variância de $s^2$ = 8,38. Determinar um intervalo de confiança de 90% para a variância da população e um intervalo de mesma confiabilidade para o desvio padrão da população.

## Solução.

Neste caso é necessário inicialmente determinar os valores da distribuição $\chi^2$, de modo, que $\chi_1^2$ tenha uma área (probabilidade) à direita igual a 95% e $\chi_2^2$ tenha uma área (probabilidade) à direita igual a 5%. Estes valores são: $\chi_1^2$ = 3,940 e $\chi_2^2$ = 18,307.

O intervalo de confiança, para a variância, será:

$$\left[\frac{(n-1)S^2}{\chi_2^2}; \frac{(n-1)S^2}{\chi_1^2}\right]$$

$$\left[\frac{(11-1).8,38}{18,307}; \frac{(11-1).8,38}{3,940}\right]$$

[4,58; 21,27]

O intervalo de confiança, para o desvio padrão, será:

$$\left[\sqrt{\frac{(n-1)S^2}{\chi_2^2}}; \sqrt{\frac{(n-1)S^2}{\chi_1^2}}\right]$$

$$\left[\sqrt{\frac{(11-1).8,38}{18,307}}; \sqrt{\frac{(11-1).8,38}{3,940}}\right]$$

$\sqrt{4,58;}\ \sqrt{21,27;}$ ]

[2,14; 4,61].

# 3. EXERCÍCIOS

**(01)** De uma população com N = 12 elementos é retirada uma amostra aleatória simples, **sem** reposição, de n = 5.

**(01.1)** Quantas são as possíveis amostras?

**(01.2)** Qual a probabilidade de cada uma destas amostras ser selecionada?

**(02)** Uma população é composta dos elementos: A, B, C, D e F.

**(02.1)** Liste todas as possíveis amostras aleatórias simples, **sem** reposição, com n = 2.

**(02.2)** Liste todas as aas, **sem** reposição, de tamanho n = 3.

**(02.3)** Determine a probabilidade de ser sorteada a amostra BC.

**(02.4)** Determine a probabilidade de ser sorteada a amostra ACD.

**(03)** A tabela, ao lado, é a distribuição de freqüências de uma amostra proveniente de determinada população.

| X | f |
|---|---|
| 1 | 40 |
| 2 | 45 |
| 3 | 8 |
| 4 | 7 |

**(03.1)** Determine o tamanho da amostra.

**(03.2)** Determine uma estimativa da média da população.

**(03.3)** Determine uma estimativa da variância da população.

**(03.4)** Determine uma estimativa da proporção de valores pares na população.

**(04)** A tabela ao lado apresenta valores amostrais.

| Elementos | X |
|---|---|
| A | 5 |
| B | 7 |
| C | 12 |
| D | 15 |
| E | 10 |

**(04.1)** Qual o tamanho da amostra?

**(04.2)** Determine uma estimativa para a média da população.

**(04.3)** Determine uma estimativa do desvio padrão populacional.

**(04.4)** Determine uma estimativa dos valores ímpares de X.

**(05)** Uma população é formada pelos elementos: A = 3, B = 6, C = 9 e D = 12.

**(05.1)** Determine os seguintes parâmetros:

**(a)** média,

**(b)** variância e

**(c)** proporção de elementos menores que 8.

**(05.2)** **(a)** Construa a distribuição amostral da média da amostra utilizando aas, **sem** reposição, de tamanho n = 2.

**(b)** Determine a expectância e a variância da distribuição amostral em (a)

**(c)** Construa a distribuição amostral da média da amostra utilizando aas, **sem** reposição, de tamanho n = 3.

**(d)** Determine a expectância e a variância da distribuição amostral em (c)

**(5.3)** **(a)** Construa a distribuição amostral da variância amostral utilizando aas, **sem** reposição, de tamanho n = 2 e determine a sua expectância.

**(b)** Utilize a correção de população finita para as variâncias obtidas em **(a)** obtendo a distribuição amostral da variância corrigida e determine sua expectância.

**(c)** Construa a distribuição amostral da variância corrigida utilizando aas, **sem** reposição, de tamanho n = 3 e determine sua expectância.

**(d)** Utilize a correção de população finita para as variâncias obtidas em **(c)** obtendo a distribuição amostral da variância corrigida e determine sua expectância.

**(5.4) (a)** Construa a distribuição amostral para o estimador da “proporção de elementos menores que 8” utilizando aas, **sem** reposição, de tamanho n = 2.

**(b)** Determine a expectância e a variância da distribuição em (a).

**(c)** Construa a distribuição amostral para o estimador da “proporção de elementos menores que 8” utilizando aas, **sem** reposição, de tamanho n = 3.

**(d)** Determine a expectância e a variância da distribuição em (c).

**(06)** Utilize os valores da amostra tabelada ao lado, extraída aleatoriamente e **sem** reposição, de uma população com N = 2000 elementos, para estimar:

| X | f |
|---|---|
| 0 \|-- 2 | 27 |
| 2 \|-- 4 | 51 |
| 4 \|-- 6 | 49 |
| 6 \|-- 8 | 48 |
| 8 \|-- 10 | 25 |

**(06.1)** A média da população.

**(06.2)** A variância da população.

**(06.3)** O percentual de elementos menores que 6.

**(06.4)** O erro amostral da média.

**(07)** De uma população com N = 4000 pessoas de uma região foi obtida uma amostra aleatória, **sem** reposição, de 400 pessoas que revelou 60 analfabetos. Estime:

**(07.1)** A proporção de analfabetos da região.

**(07.2)** O erro amostral do estimador proporção.

**(08)** Uma aas de tamanho 900 extraída de uma população bastante grande apresentou 40% de pessoas do sexo masculino. Estime o erro amostral do estimador proporção de pessoas do sexo masculino.

**(09)** Uma população tem média 500 e desvio padrão 30.

**(09.1)** Determinar a probabilidade que uma aas de 100 elementos apresentar um valor médio superior a 504,50.

**(09.2)** Calcule a probabilidade de que uma aas com n = 64 valores apresentar média entre 492,5 e 507,5.

**(09.3)** Se uma aas de n = 144 for extraída desta população, qual o percentual de médias amostrais que estarão entre 495,5 e 504,5?

**(10)** Uma população é normalmente distribuída com média 800 e desvio padrão 60.

**(10.1)** Determine a probabilidade de que uma aas de tamanho 9 apresentar média menor que 780.

**(10.2)** Calcule a probabilidade de que uma aas de tamanho n = 16 tenha média entre os valores 781,4 e 818,6.

**(10.3)** Que percentual de médias amostrais de uma amostra de tamanho n = 25 estarão no intervalo [776; 824]?

**(11)** A proporção de eleitores de um candidato é 20%.

**(11.1)** Qual a probabilidade de uma amostra aleatória simples de 100 eleitores apresentar uma proporção amostral superior a 26%?

**(11.2)** Qual a probabilidade de uma amostra aleatória simples de 400 eleitores apresentar uma proporção de eleitores do candidato entre 17% e 23%?

**(11.3)** Se a amostra aleatória for de 625 eleitores, qual a percentual de valores do estimador proporção amostral que estarão no intervalo [0,16864; 0,23136]?

**(12)** Admitindo que a probabilidade nascer um menino ou uma menina seja iguais, determine a probabilidade de que das próximas 400 crianças a nascerem:

**(12.1)** Menos de 45% sejam meninas.

**(12.2)** Mais de 54% sejam meninos.

**(13)** De uma distribuição normal com variância 2,25, obteve-se a seguinte amostra:

**27,5; 25,6; 28,2; 26,1 e 25,0**

Determinar um intervalo de confiança para a média desta população com confianças de:

**(13.1)** 95% **(13.2)** 99%

**(14)** Através de uma aas de 145 profissionais de certa região, verificou-se que o salário médio é de 8 salários mínimos (s.m.) com um desvio padrão de 1,8 s.m. A amostra também forneceu a informação de que 70% dos profissionais eram casados.

**(14.1)** Determine e interprete o intervalo de confiança de 95% para o salário médio de todos os profissionais desta região.

**(14.2)** Determine e interprete o intervalo de confiança de 99% para a proporção de profissionais casados desta região?

**(14.3)** Determine e interprete um Intervalo de Confiança de 90% para $\sigma^2$.

**(15)** A amostra apresenta os valores da variável "tamanho da família" coletados através de uma aas em uma vila popular.

**(15.1)** Determine e interprete o intervalo de confiança de 95% para o parâmetro tamanho familiar médio por domicílio da vila.

**(15.2)** Determine e interprete o intervalo de confiança de 90% para o parâmetro proporção de domicílios da vila com tamanho igual ou superior a cinco.

| X | f |
|---|---|
| 3 | 10 |
| 4 | 14 |
| 5 | 19 |
| 6 | 15 |
| 7 | 07 |

**(16)** A variância de uma população é 150. Deseja-se obter um intervalo de confiança para a média da população com uma confiabilidade de 95% e um erro máximo de 2. Quantos valores desta população devem ser retirados aleatoriamente?

**(17)** Quer-se estimar a média de uma população de variância desconhecida através de um intervalo de confiança de 95% e com erro de estimação máximo de 5 unidades. Através de uma amostra piloto de 100 valores a variância foi estimada em 400 unidades. Que tamanho deve ter a amostra final?

**(18)** Uma amostra preliminar de pessoas de uma determinada comunidade apresentou 18% de analfabetos. Com este resultado quer-se estimar a proporção de analfabetos da população com uma confiabilidade de 95% e com um erro de estimação máximo de 2,5%. Qual o tamanho da amostra a ser utilizada?

**(19)** De uma população normalmente distribuída foi extraída uma aas de n = 10 que apresentou os valores abaixo:

**4 8 12 5 7 9 10 11 6 8**

**(19.1)** Determine uma estimativa da variância populacional.

**(19.2)** Determine uma estimativa da média populacional e do correspondente erro amostral?

**(19.3)** Determine um intervalo de confiança de 95% para a média desta população.

**(20)** A tabela apresenta os valores de uma amostra retirada de uma população normal. Determine:

**(20.1)** Um intervalo de confiança de 95% para a média desta população.

**(20.2)** Um intervalo de confiança de 99% para a média desta população.

| X | f |
|---|---|
| 04 \|-- 08 | 8 |
| 08 \|-- 12 | 8 |
| 12 \|-- 16 | 6 |
| 16 \|-- 20 | 4 |

# 4. RESPOSTAS DOS EXERCÍCIOS

**(01) (1.1)** 792 **(1.2)** 1/792

**(02) (2.1)** AB AC AD AE BC BD BE CD CE DE
**(2.2)** ABC ABD ABE ACD ACE ADE BCD BCE BDE CDE
**(2.3)** 1/10 **(2.4)** 1/10

**(03) (3.1)** 100 **(3.2)** 1,82 **(3.3)** 0,73 **(3.4)** 0,52 = 52%

**(04) (4.1)** 5 **(4.2)** 9,80 **(4.3)** 3,96 **(4.4)** 0,60 = 60%

**(05) (5.1) (a)** $\mu = 7{,}50$ **(b)** $\sigma^2 = 11{,}25$ **(c)** $\pi = 0{,}50$

**(5.2)**

**(a)**

| $\bar{x}$ | 6 | 7 | 8 | 9 |
|---|---|---|---|---|
| $f(\bar{x})$ | 1/4 | 1/4 | 1/4 | 1/4 |

**(b)** $E(\bar{X}) = 7{,}50$ $V(\bar{X}) = 3{,}75$

**(c)**

| $\bar{x}$ | 4,5 | 6,0 | 7,5 | 9,0 | 10,5 |
|---|---|---|---|---|---|
| $f(\bar{x})$ | 1/6 | 1/6 | 2/6 | 1/6 | 1/6 |

**(d)** $E(\bar{X}) = 7{,}50$ $V(\bar{X}) = 1{,}25$

**(5.3)**

**(a)**

| $s^2$ | 4,5 | 18,0 | 40,5 |
|---|---|---|---|
| $f(s^2)$ | 3/6 | 2/6 | 1/6 |

$E(S^2) = 15 \neq \sigma^2$

**(b)**

| $\hat{s}^2$ | 3,375 | 13,500 | 30,375 |
|---|---|---|---|
| $f(\hat{s}^2)$ | 3/6 | 2/6 | 1/6 |

$E(\hat{S}^2) = 11{,}25 = \sigma^2$

**(c)**

| s2 | 9 | 21 |
|---|---|---|
| f(s2) | 1/2 | 1/2 |

$E(S^2) = 15 \neq \sigma^2$

**(d)**

| $\hat{s}^2$ | 9 | 17 |
|---|---|---|
| $f(\hat{s}^2)$ | 1/2 | 1/2 |

$E(\hat{S}^2) = 11{,}25 = \sigma^2$

**(5.4)**

**(a)**

| p | 0 | 0,5 | 1 |
|---|---|---|---|
| f(p) | 1/6 | 4/6 | 1/6 |

**(b)** $E(P) = 0{,}50$ $V(P) = 1/12$

**(c)**

| p | 1/3 | 2/3 |
|---|---|---|
| f(p) | 1/2 | 1/2 |

**(d)** E(P) = 0,50 V(P) = 1/36

**(06) (6.1)** $\bar{x}$ = 4,93 **(6.2)** $s^2$ = 6,1628 **(6.3)** p = 63,50% **(6.4)** 0,1666

**(07) (7.1)** 60/400 = 15% **(7.2)** 1,69%

**(08)** 1,63%

**(09) (9.1)** 6,68% **(9.2)** 95,44 **(9.3)** 92,82%

**(10) (10.1)** 15,87% **(10.2)** 78,50% **(10.3)** 95,44%

**(11) (11.1)** 6,68% **(11.2)** 86,64% **(11.3)** 95%

**(12) (12.1)** 2,28% **(12.2)** 5,48%

**(13) (13.1)** [25,17; 27,79] **(13,2)** [24,75; 28,21]

**(14) (14.1)** [7,71; 8,29] Tem-se 95% de certeza de que o salário médio de todos os profissionais da área está entre 7,71 s.m. e 8,29 s.m.

**(14.2)** [60,20%; 79,80%] Tem-se 99% de confiança de que a percentagem de profissionais da área que são casados esteja entre 60,20% e 79,80%.

**(14.3)** [2,70; 3,98]. Tem-se 90% de confiança de que o valor da variância populacional pertença a este intervalo.

**(15) (15.1)** [4,62; 5,22] Tem-se 95% de confiança de que o valor médio do tamanho familiar da vila esteja entre 4,62 e 5,22 membros.

**(15.2)** [53,23%; 72,93%] Há 90 de certeza de que o percentual de famílias com 5 ou mais membros esteja entre 53,23% e 72,93%.

**(16)** n = 145

**(17)** n = 62, como a amostra piloto utilizada foi de n = 100 é mais confiável ficar com a amostra piloto.

**(18)** n = 908

**(19) (19.1)** 6,67 **(19.2)** 8 e 0,82 **(19.3)** [6,15; 9,85]

**(20) (20.1)** [9,19; 12,65] **(20.2)** [8,58; 13,26]

# 5. REFERÊNCIAS

[BUS86] BUSSAB, Wilton O, MORETTIN, Pedro A. *Estatística Básica*. 3. ed. São Paulo, Atual, 1986.

[HOF80] HOFFMAN, Rodolfo. *Estatística para Economistas*. São Paulo. Livraria Pioneira Editora, 1980.

[NET74] NETO, Pedro Luiz de Oliveira Costa. *Estatística*. São Paulo, Edgard Blücher, 1977.

[NET74] NETO, Pedro Luiz de Oliveira Costa, CYMBALISTA, Melvin. *Probabilidades: resumos teóricos, exercícios resolvidos, exercícios propostos*. São Paulo, Edgard Blücher, 1974.

[MAS90] MASON, Robert D., DOUGLAS, Lind A. *Statistical Techniques in Business And Economics*. IRWIN, Boston, 1990.

[MEY78] MEYER, Paul L. *Probabilidade: aplicações à Estatística.* Tradução do Prof. Ruy C. B. Lourenço Filho. Rio de Janeiro, Livros Técnicos e Científicos Editora S.A., 1978

[STE81] STEVENSON, William J. *Estatística Aplicada à Administração*. São Paulo. Editora Harbra, 1981.

[WON85] WONNACOTT, Ronald J., WONNACOTT, Thomas. Fundamentos de Estatística. Rio de Janeiro. Livros Técnicos e Científicos Editora S. A., 1985.